

DOCUMENT RESUME

ED 174 630

TM 008 478

AUTHOR Sigel, Irving E.
 TITLE A Developmental Perspective in Evaluating Educational Programs.
 PUB DATE Mar 78
 NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF01/PC01 Plus Postage.
 DESCRIPTORS *Child Development; Cognitive Measurement; *Developmental Psychology; Developmental Stages; Educational Research; Educational Testing; Evaluation Methods; Longitudinal Studies; Preschool Education; Preschool Programs; *Program Evaluation; Psychometrics; *Reliability; Research Needs; *Student School Relationship; Test Construction; *Testing Problems; Test Interpretation; Validity

ABSTRACT

This paper provides a theoretical discussion of educational program evaluation. Psychometric theory and developmental psychology are compared as they pertain to the testing of children. The nature of change in childhood makes it necessary to examine the assumptions and goals related to the testing of children as a means of evaluating educational programs. Specific issues considered include: test validity and reliability, test content, structure or subject variables, the role of environmental situation in children's behavior or test performance, and longitudinal issues. The evaluation of a preschool program is described to illustrate various evaluation issues. (RD)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

A Developmental Perspective in Evaluating Educational Programs

Irving E. Sigel

Educational Testing Service

Princeton, New Jersey 08540

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

I. Sigel

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC) AND
USERS OF THE ERIC SYSTEM."

Paper presented at the meeting of the American Educational Research Association, Toronto, Canada, March 1978.

A Developmental Perspective in Evaluating Educational Programs

Irving E. Sigel

Educational Testing Service

Great progress has been made in the field of educational measurement. There have been theoretical and methodological advances in psychometrics with a move away from the simplistic concepts of measurements to more sophisticated conceptualizations of the issues and problems in the light of increasing psychological knowledge of human development. Unfortunately, despite the advancements that have occurred within the domains of psychometrics and developmental psychology, there has not been a joining of the two forces in the development of the evaluation process of education programs.

The reason for this concern is that, in recent years, considerable attention has been directed toward evaluating the effectiveness of educational programs in enhancing the quality of education. The consequences of the results of such studies, not only for theory building regarding the educational process, but equally important for educational planning, and social policy, are considerable. If public support for education is contingent on the results of psychometric evaluations, then the practical significance is even more dramatic.

Acceptance of psychometric results as valid indicators of educational progress rests on the proposition that the instruments employed to assess the program outcomes are accurate, valid procedures. If the measures are valid, then it would follow that the lack of expected progress among children, for example, is due either to the inadequacy of the educational program or to the nature of the children. On the other hand, it may well be that the

educational program is adequate and the children do profit, but that the assessment procedures are irrelevant to school experiences.

The concept of validity in itself poses some problems since one must deal with four types of validity: predictive, concurrent, construct and content validity (Cronbach & Meehl, 1967). Predictive and concurrent validity differ in the conceptualization of the items involved. The differences and similarities among these conceptions of validity have been dealt with elsewhere (Messick, 1974). For present purposes, let it suffice to say that common to all of them is the idea that the instrument employed does in fact measure that characteristic which it purports to measure.

The concept of validity in this discussion is no different from the one employed in any measurement context. The basic problem is the definition of the variables that are to be assessed, and my belief is that for children some of the measures are unique. This leads us directly to the question of reliability, classically defined as consistency and coherence, which also poses a unique problem for child development researchers.

The two tests of reliability most commonly reported are split half and test-retest. For the developing organism a unique measurement problem exists; for consistency within a task, fatigue and boredom have a great probability of occurring among children reducing the intercorrelation among tasks. Test-retest, on the other hand, poses the problem of how far apart tests can be given, since the child's development is an ongoing rapid process with the rate of change not always regular or predictable. Thus, the time span between testings is a critical consideration.

Meeting these traditional measurement objectives results in violating the psychological principle related to a conceptualization of the child as a

dynamic, living organism locked into a system of transformational change.

Measurement of adults presumes that the characteristic being measured is stable and that the relationship obtained among variables is independent of temporal relationships. For a developmental psychologist, this perspective is inappropriate, because a static state is a nonreal phenomenon at any age. Change permeates all living matter, even adults, though change in them is perhaps at a slower rate. For children in whom change is dramatic and often apparent but unpredictable, we are faced with a ticklish measurement problem when using test-retest procedures. How long a temporal interval should be used? If we give the child a test at Period 1 and retest him at Period 2, what are the parameters of this time span--a day, a week, a month--to be considered as a test of reliability? Binet implies a six month interval judging by the mental age levels prior to age 5. After age 5, a year is used as the scored time interval. This implies that after 5 years of age, change is less dramatic. After what time period do we consider it no longer an issue of reliability of the test, but the stability of the behavior in question? The distinction between reliability of the measure and stability of the individual becomes somewhat clouded and can only be judged on the basis of the knowledge we have of the developmental process; this then becomes a validity issue. Validity and reliability begin to blend because the consistency over time may be reliable, but one needs to question whether it is, in fact, consistency. If relative position of individuals within a group is maintained on a retest, but the scores are higher, is the test reliable and also valid? Decisions about the time span have to be made in educational evaluation. In every case, we are dealing with a longitudinal study--whether the design calls for a pre-post assessment over a semester or the academic year, the basic design

must be conceptualized as longitudinal. The only nonlongitudinal assessment is a one time comparison across groups matched on age. Thus, the issue we must come to terms with is how to conceptualize and appropriately establish the methodology for short- or long-term longitudinal studies--essentially developmental.

Development refers to transformational changes in competence levels. Since in educational research we are dealing with children, we must integrate a conception of change in our efforts. Change, from my perspective, is not just a simple accretion of skills or knowledge. Rather, change involves qualitative differences--it is not just knowing that $5 \times 5 = 25$, but what principle is involved, it is moving from concrete to abstract thought, from egocentric perspectives to sociocentric perspectives (Piaget, 1972).

The conceptualization of development must permeate all phases of the evaluational planning. Thus rather than elaborate on a discussion of it now, let us leave it and return to it during the course of this discussion.

Is Treatment Exposure to Educational Programs Equal for all Participants?

A critical issue is, does every individual in the group (assuming the education is viewed as a treatment condition) receive equal treatment or is subject to the same treatment in the program?

There is no doubt that in a group of 30, 50, or 100 eight-year-old children there will be variations in developmental maturity level relative to some characteristic. This is what would account for variability in initial performance. The educational program, when introduced, will be responded to and coped with differentially--in part a function of these differences. The degree to which the child assimilate the new treatment (educational program) is an unknown. In fact, it is rarely dealt with as a relevant issue. Yet when change is evaluated and causal connections are sought, a wide gap in understanding must surely remain.

Thus, it is necessary in evaluation research to monitor the program ascertaining the participation of each member in the experiment, and assess the initial status of the child developmentally--in particular, relevant to the educational program involved.

Let me illustrate how we, in a preschool program, attempted to fulfill these two conditions.

- (1) Educational objectives in terms of (1) processes involved, (b) role of content, (c) support system organization for treatment were clearly defined.
- (2) The treatment condition (teacher strategy, materials, etc.) was clearly articulated.
- (3) Creation of observation schedule to monitor teachers execution of the program. A time sampling observation of teachers, children, and setting was done.
- (4) Feedback to teachers regarding fulfilling program objectives was regularly carried out.
- (5) Assessment procedures with children pre-post on measures conceptually related to program theory and educational objectives. Thus, tests that tapped those processes were hypothesized to be affected. Lest this be thought of as teaching to test, let me reassure you this is not the case. Aside from the usual safeguards regarding teacher knowing what tests are used, our measures are Piaget-type tasks which are assumed to tap processes, e.g., memory, imagery, anticipation, processes that are engaged in a variety of school contexts. But there is no one-to-one correspondence between classroom and assessment.

These procedures will, I believe, provide greater assurance that factors antecedent to outcomes can be identified, and some degree of certainty regarding educational antecedents of outcomes will be achieved.

Developmental Change and Test Performance

Let me demonstrate the issue of change in the context of a research project which was a longitudinal intervention study begun with children at age two and carried on until they were age five (Sigel, Secrist, & Foreman, 1973). We devised measurement procedures to evaluate program effectiveness; the theory of the program dictated the areas of measurement. Thus, we evaluated performance in certain cognitive and social domains. Since the program was longitudinal in nature, we had to find tests which could be used at different age levels (2, 3, 4, and 5 years), or if not, find instruments that presumably would measure the same variable at each time point.

The question is: Does a test (controlled for difficulty, of course) measure the same variable at each point in time in the course of development? If, for example, the child is given a sorting task with familiar objects, does it measure classification ability each time it is used or does it measure different things at different times? To check this we analyzed relationships among our battery of tests at each of the time periods, and indeed the tests did not cluster similarly each time. We could not attribute this to practice effects, since the clustering did not appear to include tests that could be improved on by repeated measurement (Sigel et al., 1973). Why would a test load on one factor at one point in time and on another factor at another point in time? The explanation may reside in one of two sets of elements in the testing situation: test content and structure or subject variables which may be particularly sensitive to developmental changes.

The Test Situation and Subject Variables

The format of a test may well influence performance. For example, the use of three-dimensional objects in contrast to two-dimensional objects might

make a difference (Sigel & McBane, 1967). Structure of the questions, length of the test, and interrogations are among the types of factors that may influence the reliability and validity of a child's test performance.

Perhaps more important for our consideration are the subject variables. Since we are working with relatively rapidly changing organisms, it may be that the measure in question taps a particular kind of competence that is in an emergent state and when manifested, functions differently, e.g., space relationships or language items. The child may show minimal ability to handle such a task at the initial testing, but as time passes he may develop a competency in an area which has broader impact, influencing not only the particular area but related areas. In other words, cognitive growth involves more than a single area of development.

Such findings may also be due to familiarity with the task or increasing maturity, where the child develops a different perspective regarding the task, or where the solution to the problem for him may involve processes not previously available.

Granting this as the issue, let us turn to another set of problems that also focus on the subject variables and are related to the fact that in testing very young children we are working with a socially inexperienced organism that is very sensitive to changes in the environment. The data from our project showed that children who were most socially expressive and outgoing seemed to perform better than those who were quiet and reserved. The fact that the former were mostly girls suggested that there was a test situation by sex interaction (Sigel et al, 1973). This finding is perhaps consonant with the usual report that girls are more expressive and more relationship-oriented. The question is, "Which comes first?" Is the girl

competent and her social behavior expressive of her personal security, or do social behaviors facilitate test performance since they call attention to the tester, thereby resulting in better comprehension of what the test tasks involve? Whatever the reasons, there is clear-cut evidence in our results and in the results of other investigators that the "personal-social" relationship of the child and the examiner influences test results.

Before proceeding further, I want to dwell on one issue that is particularly crucial: namely, the longitudinal aspect of our test results. Previously, it was noted that we found test results that did not correlate in a similar way at different ages.

Stability, Continuity or Transformation

This phenomenon highlights further the need to conceptualize developmental change. In personality theory there has been some discussion of stability versus continuity of traits. Are traits consistent over time, which means that they function and are identifiable in the same form from one time epoch to another; or are they similar in form but different in meaning from one time epoch to another; or are they different in form and in meaning from one epoch to another; or are they similar in form and in meaning from one time epoch to another?

Each of these is possible. For example, does aggression serve the same psychological function when manifested by a five- or ten-year-old? Does the same act appearing at age five differ in significance when expressed at age 10, since the child at this older age knows different things about social behavior? Do we call them the same aggressive behaviors? The problem of measurement in the personal-social domain is that the behaviors defined at one age level can be similarly defined at another age level, but the function

of that behavior may vary from one point in time to another.

The problem in the cognitive area is similar, although in some respects it is a bit more complex. At one age level, for example, there is no evidence of formal reasoning and at another the child is logical--is that a discontinuous change or a continuous change?

The issues of continuity, stability, or transformation form the crucial challenges to developmental measurement. Results from evaluation of our program tend to suggest that similarities in form do not mean similarities in function (Sigel et al., 1973). This is to say, in terms of traits or abilities, the appearance of something that looks like something else on a manifest level may in fact not function in an identical way. For example, memory for digits for the very young child may be an index of comprehension of ordering numbers, whether arbitrary or not, while at an older level it may appear as a rote memory task. On a draw-a-person task some children at young age levels may in fact draw a conception of man, while the older child may well draw an idealized image, not really a reflection of the child's own perceptions of reality.

A few efforts at conceptualizing these issues have been made by Kagan (1969) and Emmerich (1966), and by Wohlwill in his recent eloquent book on methodology (Wohlwill, 1973). The point is that in the field of measurement these issues are just as significant as they are in the field of theory and we shall have to address ourselves to them. As Wohlwill has so aptly put it:

...the field of stability and continuity is one strewn with roadblocks of the investigator's own making, arising from the purely correlational sense in which the terms are typically taken, combined with the all too human tendency to lose sight

of the limitations in discussion of the subject. (It is interesting to note that the same difficulty is encountered in the field of educational measurement in the interpretation of test-retest correlations of a single test...) (Wohlwill, 1973, p. 373).

What Wohlwill argues for is a clearer conceptualization of continuity.

Continuity poses particular problems for evaluation. For some phenomena continuity in the sense of incremental gains, e.g., better skill in playing tennis or doing arithmetic, suggests tests that are cumulative in that the items are of the same class but get harder. For other phenomena continuity is conceptualized as correlated with subsequent behavior, but the correlated behaviors are not expressed in the same form; that is, a transformation has taken place. This type of apparent discontinuity with underlying continuity would suggest different types of items and tasks.

It would be argued that both types of continuity exist, depending on what types of tasks are involved. Clarification of the concept of continuity should influence the types of measures used. This suggests, then, that it will behoove evaluators in assessing programs to consider continuity, especially when they are concerned with change over time, a usual programmatic requirement. I would suggest that for a behavioristically-oriented program, a cumulative perspective would be more likely; for Piagetian-oriented programs, transformational perspectives would prevail.

Same Person--Different Settings

If we use the argument that test performance is influenced by the setting and the format in which the test is given, then we must so to speak write this into our evaluation equation. In general, we have three options: group testing, one-to-one testing, individualized evaluation, e.g.,

program instruction or other individualized work. Each of these situations creates its own set of sources of statistical and psychological error.

Thus, evaluation is perforce an approximation. The pros and cons of each of these procedures have oft been repeated. I mention them to fold these types of issues into my general argument regarding the factors that must be considered in evaluating the quality of the evaluation.

People are products of their environment and act differently in varying settings. Our task is to untangle this mess without violating the integrity of the mess! Working with a child in a nursery school or in a clinic, one is immediately brought up short regarding how complex the child is in the here and now--how unpredictable and yet how predictable he is.

The determinants of the child's complexity reside in his inheritance and in his experience interacting within the context of a particular setting. We are in fact dealing with a behavioral field, and the forces within this field impinge on the individual. Lewin (1951) said it long ago; Barker (1963) and Gump (1969) have said it; and Koffka's (1935) distant voice said it loud and strong, but the behaviorists in the U.S. never really heard it. Our task is somehow to understand the significance of the various settings in which a child lives and then attempt to get at the interactions within these settings, examining how they influence the kind and the quality of ongoing behavior. Willems and Willems (1965) have shown quite clearly that the data obtained about a child do depend on the method used to gather the data. Is this not reasonable? After all, is it not your experience that you are not the same "you" in the bathtub as the "you" in the classroom; that you are not the same "you" at a cocktail party as in a library. If we agree that we manifest differences from one setting to another, is it unreasonable to suppose that

the child also changes behavior from one setting to another?

Our testing rationale does not take such differences into account. The tests we use are presumed to transcend the specific situation, have general predictive value, and presume to predict to situations in general. Let me ask why do we want to know a child's IQ? To know how smart he is? Well and good, but how smart in what situation--on the baseball field, in the classroom, at a tea party, while working a mechanical puzzle? No, we do not specify the situation for which we are making the prediction. Validity might be enhanced if we consider the setting and further specify the relevant dimensions of the setting. If the setting, the nature of the items, and the social situation each contribute to the reduction of validity, then it is incumbent upon us to discover ways to cope with this problem.

Longitudinal Issues

The school system is a socializing experience that involves the child for about 13 years of its life. Embedded in the school setting, the child is passed through a variety of steps, each of which is defined as a prelude to the next. Thus it is reasonable to assume that an interconnectedness exists in the educational enterprise. The test of the interconnectedness is to evaluate performance at each level relative to its antecedents. Or to put it another way, is there any predictability for performance at one time point to another.

Testing the interdependence of performance across time points would seem to be a worthwhile effort because it would validate the integral nature of the educational effort; it should yield important data regarding program planning in defining prerequisites. Theoretically it could provide important information on such topics as developmental structure of intellect.

Finally, proper evaluation of change should enable us to identify with greater specificity factors influencing course of growth.

Unfortunately, little has been done to exploit this natural experiment. Rather, conventional tests are used, and a longitudinal paradigm may be adhered to only in the sense of testing children at varying age levels. Yet to determine the impact of educational experience in an accumulative way longitudinal assessment is necessary. In the main, the challenge to educational evaluation is not in the elegance of the research design or data analysis. We know a lot about these. Rather, I find a sterility in much of the research because of a theoretical disregard of a critical aspect in the effort, namely, the developing child functioning in an ongoing changing system. Unless the dynamics are built into the evaluation system research efforts in that direction may tell us a lot about an unimportant aspect of the educational enterprise.

It might then be asked, "Of what value is this assessment information if it is not predictive of the child's behavior in other contexts?" The predictive power is variable. Some children will be consistent: that is, high correlations will be obtained between settings; whereas for other children this will not be the case. The task is to discover how we can know this in advance. Bem and Allen (1974), working with college students, asked individuals whether or not they considered themselves consistent, and this was a fruitful procedure. We cannot ask these questions of young children, and therefore, have to find other methods to determine consistency.

One method is to try to get the information on the same children's behavior in different settings and from different informants such as teachers and parents. Asking teachers to rate children on behaviors similar to those

assessed in the test situation might be appropriate. Apparently, in the past the tendency was to disregard these ratings because of low correlations with test behavior. This may have been an error, since each set of data may have been accurate. In one setting the child displays a set of competencies, and in another context he might not. The situational demand factors may make the difference, so certain behaviors may not appear or may appear in a different form. What we may have done in the past was to overlook the situational variables, thereby assuming low correlations were due to inadequate raters or inadequate methods. The actual reason for low correlations might have been an inadequate conception of the variability in an individual's behavior relative to the setting.

This argument would speak first to a reconsideration of ratings or other types of assessment of children in different settings, and second to the issue that data from any situation are limited not because of the "fault" of the examiner, the child, or the instruments, but rather because the type of setting places constraints on the individual. If one accepts this argument, then there is need for the delineation of details of the situation to discover what the salient factors are. There is also need to construct other procedures which will flesh out findings based on any one particular situation.

Defining the variables in the one-to-one situation might contribute to our understanding of the interactional nature of the setting, that is, the child, the tasks, and the interrogator. Some investigators have already proceeded in this direction (Schoggen & Schoggen, 1968). As one is concerned about the nature of the task, so one should be concerned about the nature of the setting. For researchers, analysis of components of the setting is vital. But since we can't wait for the results of needed studies in this area and

must make some action decisions now, I would urge that methods be created, however crude, to attend to the description of the situation in the following terms: physical setting, task, type of interaction (e.g., interview, conversation, didactic), and length of time involved, among others. The examiner should attend to all of those classes of behavior which are deemed theoretically relevant to the performance, and which are quantifiable.

References

- Barker, R. G. (Ed.) The stream of behavior. New York: Appleton-Century-Crofts, 1963.
- Bem, D. & Allen, A. On predicting some of the people some of the time. Psychological Review, 1974, 81, 506-520.
- Cronbach, L. J. & Meehl, P. E. Construct validity in psychological tests. In D. M. Jackson, & S. Messich (Eds.), Problems in human assessment. New York: McGraw-Hill, 1967.
- Emmerich, W. Continuity and stability in early social development: II. Teacher ratings. Child Development, 1966, 37, 17-27.
- Gump, P. V. Intra-setting analysis. The third grade classroom as a special but instructive case. In E. P. Willems and H. L. Raush (Eds.), Naturalistic viewpoints in psychological research. New York: Holt, Rinehart, & Winston, 1969.
- Kagan, J. The three faces of continuity in human development. In D. A. Goslin (Ed.), Handbook of socialization theory and research. Chicago: Rand McNally, 1969.
- Koffka, K. Principles of Gestalt psychology. New York: Harcourt Brace & Co., 1935.
- Lewin, K. Field theory in social science. New York: Harper & Row, 1951.
- Messick, S. The standard problem: Meaning and values in measurement and evaluation (ETS RB-74-44). Princeton, N.J.: Educational Testing Service, 1974.
- Piaget, J. Psychology of intelligence. Totowa, N.J.: Littlefield, Adams & Company, 1972.

- Schoggen, P. & Schoggen, M. Ecological approaches to the study of preschool environments. Demonstration and Research Center for Early Education, George Peabody College for Teachers, Nashville, Tennessee, 1968.
- Sigel, I. E., & McBane, B. Cognitive competence and level of symbolization among five year old children. In J. Hellmuth (Ed.), The disadvantaged child. Vol. 1. Seattle, Washington: Special Child Publications, 1967.
- Sigel, I. E., Secrist, A., & Foreman, G. Psycho-educational intervention beginning at age two: Reflections and outcomes. In J. Stanley (Ed.). Compensatory education for children age two to eight. Baltimore, Md.: The Johns Hopkins University Press, 1973.
- Willem, E. P. & Willem, G. J. Comparative validity of data yielded by three methods. Merrill-Palmer Quarterly, 1965, 11, 65-71.
- Wohlwill, J. The study of behavioral development. New York: Academic Press, 1973.

Footnote

¹Portions of this paper were reported in: Sigel, I. E. The search for validity or the evaluator's nightmare. In Richard A. Weinberg and Shirley C. Moore (Eds.), Evaluation of educational programs for young children: The Minnesota Round Table on Early Childhood Education II. Washington, D.C.: The Child Development Associate Consortium, 1975. Pp. 53-66.